Chapter 13
Decoding Algorithmic Bias: Definitions, Sources, and Mitigation Strategies

Ozgur Aksoy
Istanbul Universitesi, Turkey

ABSTRACT

Predictive algorithms are increasingly used to assist decision-making for efficiency gains. However, it is essential to acknowledge that algorithms can mirror systemic biases in their predictions in a way that favors certain groups over others, even if they are immune to cognitive biases. The notion of algorithms generating unfair predictions is referred to as "algorithmic bias." Addressing cognitive biases in humans might not always be an effective solution to mitigate algorithmic bias. Therefore, it is essential to understand when and how quantitative technical mitigation methods can address this issue. This chapter explores the fundamental concepts of algorithmic bias, its sources, and technical mitigation strategies. In a world where humans and AI are intertwined, it is our responsibility to ensure a fair digital future. Addressing algorithmic bias is critical to achieving this goal.

"What challenges us is to ensure that none should enjoy lesser rights; and none tormented because they are born different, hold contrary political views or pray to God in a different manner." ~Nelson Mandela, 1995.

INTRODUCTION

Predictive algorithms are increasingly used to assist decision-making, such as loan approvals and resource allocation for patients with complex health needs. For instance, in the banking sector, an algorithm can scrutinize a credit card applicant's past transaction to determine whether they pose a risk of non-payment. This enables loan officers to process many credit card applications quickly and efficiently. Similarly,

DOI: 10.4018/979-8-3693-1766-2.ch013

algorithms in the healthcare industry can analyze past data to propose personalized treatment plans for patients at high risk of contracting multiple conditions. By examining data from electronic health records, algorithms can provide healthcare professionals with more comprehensive insights for making patient care decisions. Implementing predictive algorithms in resource allocation decisions leads to significant efficiency gains.

While algorithms are immune to cognitive biases, they can still generate systematic and repeatable errors, leading to unfair predictions. This phenomenon is known as "algorithmic bias," a term that describes a situation where a computer system inadvertently creates an unfair advantage for certain groups through its algorithmic predictions.

Algorithmic bias is primarily because the historical data used to train these algorithms often mirrors the accumulated biases in human decision-making. Even when presented with identical inputs, humans often make divergent decisions across multiple instances (Kahneman et al., 2016). For example, Bertrand and Mullainathan (2004) found that interview callbacks were 50% more likely for applicants with White-sounding names than those with Black-sounding names. Similarly, Van Such et al. (2017) discovered that 21% of cases had distinctly different diagnoses between referral and final diagnoses in a sample of 286 patients. Over time, these variations in decision-making can accumulate, leading to systemic biases that can affect our institutions and relationships (Vaught & Castagno, 2008) and the decision-making systems we create.

Research in various fields, such as human resources (Feldman et al., 2014), healthcare (Ahsen et al., 2018), and criminal justice (Chouldechova, 2017; Kasy & Abebe, 2021), has highlighted the prevalence of algorithmic bias. Lambrecht and Tucker (2019) found that an algorithm designed to deliver gender-neutral job ads in science, technology, engineering, and math disproportionately reached fewer women than men. Angwin and Larson (2016) evaluated a recidivism prediction instrument called COMPAS, determining bias toward Black defendants. The authors discovered that non-recidivating Black defendants were roughly twice as likely to be high-risk as White defendants and Black recidivists had half the possibility of being low risk as Whites. Caliskan et al. (2017) showed that, in a natural language processing algorithm, female words (e.g., "woman," "girl") are more associated than male words with the arts than mathematics. They also found that names linked with Black people were substantially more connected with unpleasant terms than pleasant terms when compared to White ones.

As decision-makers increasingly rely on algorithms to make critical decisions, it becomes inevitable that algorithmic bias will seep into artificial intelligence (AI) systems. While one might recommend avoiding algorithms altogether to prevent bias or focus only on biases that arise in human decision-making, these approaches will not address all biases in this digital era. As AI systems advance to the point where generative AI systems recommend treatments for patients or make investment decisions for consumers, it is necessary to find ways to address and mitigate biases in these systems to ensure they are fair for all users. To do this, thoroughly understanding the factors contributing to algorithmic bias is essential.

Algorithmic bias may arise from various sources, such as biased data sets or flawed algorithms. Identifying the exact source of algorithmic bias and developing appropriate strategies to mitigate it effectively is essential. For instance, if the bias stems from a biased data set, a possible mitigation strategy would be to ensure that the data set is more diverse and representative. Similarly, if the algorithm's mechanism causes disparities in algorithmic predictions, code updates might help eliminate the bias.

Achieving a balance between reducing algorithmic bias and maximizing efficiency through technical solutions can be challenging. Nevertheless, it provides valuable opportunities that are not possible in human-only systems. If algorithmic bias arises from the cognitive biases reflected in the historical